Name: Achyut Shukla Batch: ML - 1 PRN: 20070122005

॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)

(Established under section 3 of the UGC Act 1956)

Re - accredited by NAAC with 'A' Grade

Founder: Prof.Dr. S. B. Mujumdar, M.Sc., Ph.D. (Awarded Padma Bhushan and Padma Shri by President of India)

Lab Assignment —10

Aim :

Implement Hierarchical Clustering Analysis on a Dataset.

PART — A

Hierarchical Clustering

Hierarchical Cluster Analysis, often simply referred to as hierarchical clustering, is a popular method in data analysis and statistics used to group similar data points or objects into clusters. It is an unsupervised machine learning technique that aims to discover the underlying structure within a dataset without prior knowledge of the number of clusters.

Hierarchical clustering can be visualized as a tree-like structure, known as a dendrogram, where data points are progressively grouped together based on their similarities. The key idea is to start with each data point as its own cluster and iteratively merge the most similar clusters until all data points belong to a single cluster or a predefined number of clusters.

There are two primary approaches to hierarchical clustering:

1. Agglomerative (Bottom-Up) Hierarchical Clustering:
- Initially, each data point is treated as a separate cluster.
- At each step, the two closest clusters are merged into a single cluster.
- This process continues until all data points are in one cluster or the desired number of clusters is reached.
- The result is a dendrogram that illustrates the merging process and the hierarchical relationships between data points.

2. Divisive (Top-Down) Hierarchical Clustering:
- Initially, all data points are considered to belong to a single cluster.
- At each step, the cluster is divided into two subclusters that are less similar to each other.
- This process continues until each data point is in its own cluster or the desired number of clusters is reached. - The dendrogram can also be created, but in this case, it shows the division of clusters.

The choice between agglomerative and divisive clustering depends on the specific problem and the nature of the data. Agglomerative clustering is more commonly used and is often the preferred method in practice.

Hierarchical clustering can be used with various distance metrics to measure the dissimilarity or similarity between data points. Some commonly used distance metrics include Euclidean distance, Manhattan distance, and correlation distance, among others.

Advantages of hierarchical cluster analysis include:

1. Hierarchical Structure: The dendrogram provides a clear visual representation of the hierarchy of clusters, making it easy to interpret and identify meaningful subgroups within the data.

2. No Need for a Prespecified Number of Clusters: Hierarchical clustering doesn't require you to specify the number of clusters in advance, making it a flexible approach.

3. Agglomeration and Divergence: It can reveal both the merging and splitting of clusters, allowing for fine-grained exploration of data structure.

4. Robustness: Hierarchical clustering is less sensitive to initial conditions compared to some other clustering methods.

However, hierarchical clustering has some limitations as well:

1. Scalability: It can be computationally intensive, especially with large datasets, as the time complexity is often quadratic or worse.

2. Lack of Objectivity: The choice of distance metric and linkage method (how to calculate the similarity between clusters) can significantly impact the results, and there's no universal rule to guide their selection.

3. Interpretation Challenges: Interpreting the dendrogram to determine the optimal number of clusters can be subjective and context-dependent.

In summary, hierarchical cluster analysis is a valuable tool for uncovering patterns and structure within datasets when you have no prior information about the number of clusters. Its ability to create a hierarchical representation of clusters provides insights into the relationships among data points, but careful consideration of distance metrics and linkage methods is essential for meaningful results.

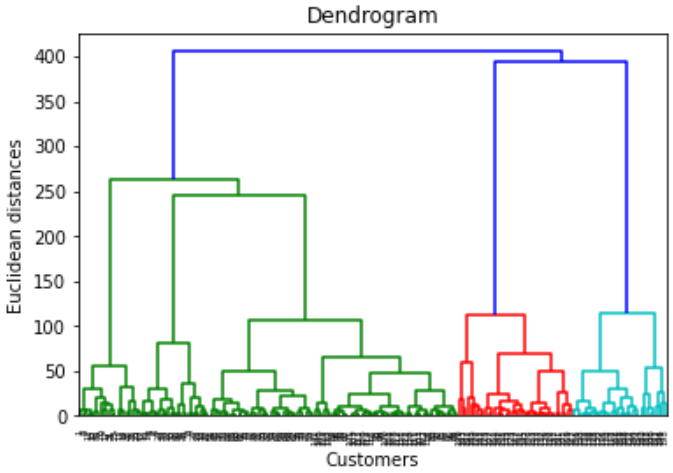
PART — B

Experiment:

Implement HCA.

Linear Regression Output:

1. Finding Optimum No of Clusters



2. Visualizing the Clusters

Inference Discussion

We have successfully implemented Hierarchical Clustering and Visualized the Clusters.